# Determining Expert Profiles
## (With an Application to Expert Finding)

**Krisztian Balog** and Maarten de Rijke

ISLA, University of Amsterdam

## Motivation

- Searching an organization's document repositories
- Finding the *right* person
  - expert on the topic
  - people you'd contact with questions on the topic
- From retrieving documents to retrieving objects

## Related work

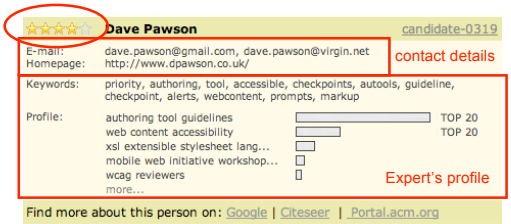
- Expert finding
  - *"Who are the experts on topic X?"*
  - Introduced at TREC in 2005
  - Given a query, return a ranked list of person names in response
- Problems:
  - Desired output should be more than a ranked list of person names
  - Context and evidence to help users

## Expert's profile

- Reverse of expert finding: *"What does expert X know?"*
- Topical profile
  - description of the areas in which she is an expert
- Social profile
  - description of her collaboration environment

## The picture

level of expertise (given the query)




*query: authoring tools*

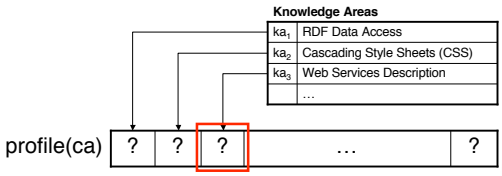
## Outline

- Introduction
- Topical profiles
  - Formal definition
  - Algorithms
  - Evaluation
- An application to expert finding
- Conclusions
- Further work

## Topical profile

Record of areas of skills and knowledge and the level of 'competency' in each.

profile(ca) = <score(ca,ka$_1$), … , score(ca,ka$_n$)>




How to estimate score(ca,ka)?

## Baseline

- Use an existing* expert finding method
  - **p(calq)** = probability of candidate ca being an expert given topic q
  - **rank(ca,q)** = position of ca on the ranked list of candidates given topic q
  - use knowledge area as the topic (q=ka)

| | |
|---|---|
| Probability baseline | score(ca,ka) = p(calka) |
| Rank baseline | score(ca,ka) = 1/rank(ca,ka) |

* K.Balog, L.Azzopardi and M. de Rijke. Formal Models for Expert Finding in Enterprise Corpora. In: SIGIR 2006

## Method 1

- find documents that are relevant to the knowledge area
- sum up the relevance of those that are associated with the person

$$score(ca, ka) = \sum_{d \in D_{ka}} relevance(d, ka) A(d, ca)$$

$D_{ka}$: set of documents relevant to *ka*; $relevance(d, ka)$: relevance of doc *d* given the topic *ka*; $A(d, ca)$: 1 if doc *d* is associated with the person *ca*, 0 otherwise

## Method 2

- represent both the knowledge area and the candidate as a set of keywords
- ratio of co-occurring keywords is regarded as being the person's competence
- each document *d* is represented as a set of keywords: KW(d)
  - keywords are extracted using TF-IDF

## Method 2 cont'd.

- Represent knowledge area as a set of keywords:

$$KW_{ka} = \bigcup_{d \in D_{ka}} KW(d)$$

- Represent individual as a set of keywords:

$$KW_{ca} = \bigcup_{d \in D, A(d,ca)=1} KW(d)$$

- Estimate the person's competence with the ratio of co-occurring keywords:

$$score(ca, ka) = |KW_{ka} \cap KW_{ca}| / |KW_{ka}|$$

## Filtering

- A knowledge area can be part of the candidate's profile, if the person is among the top *f* ranked experts on that field
- Rank experts, using the profile scores
- Use these results to refine the output of the profiling method

$$score'(ca, ka) = \begin{cases} score(ca, ka). & \text{if } |\{ca' | score(ca', ka) < score(ca, ka)\}| < f \\ 0, & \text{otherwise} \end{cases}$$

## Evaluation

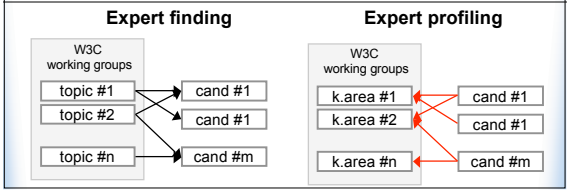
- Is “inverted expert finding” a viable solution?
- How do Method 1 and Method 2 perform?
- What is the impact of filtering?

## Evaluation (2)

- TREC Enterprise 2005 platform
- W3C collection
  - Mixture of document types crawled from w3c.org (www, wikis, e-mail lists archive, etc.)
  - 330.000 documents, 5.7 GB
- List of 1092 candidate experts
  - unique ID, name, e-mail address(es)
- 50 topics, and relevance judgments

## Creating topics and relevance judgements

- Utilize: TREC 2005 topics are names of W3C working groups
  - Use working group names as knowledge areas
  - A knowledge area is part of a person’s profile, if the person is member of the corresponding working group




## Results

| Method | MAP | MRR |
|---|---|---|
| Baseline (probability) | 0.320 | 0.397 |
| Baseline (rank) | 0.203 | 0.244 |
| Method 1 | 0.407 | 0.503 |
| Method 2 | 0.397 | 0.486 |
| Method 1 + filtering (f=15) | 0.408 | 0.649 |
| Method 2 + filtering (f=150) | 0.383 | 0.511 |

- Both Method 1 and 2 outperform baseline
- Filtering: early precision echancing effect

## Outline

- Introduction
- Topical profiles
  - Formal definition
  - Algorithms
  - Evaluation
- An application to expert finding
- Conclusions
- Further work

## An Application to Expert Finding

- Use profiles to improve on expert finding
- If a knowledge area ranks low on a person’s profile => push the candidate down on the list of experts
- Combine rankings

$$rank'_{EF}(ca, ka) =$$
$$(A) \quad = \quad \frac{1}{rank_{EF}(ca,ka)} \frac{1}{rank_{PR}(ca,ka)}$$
$$(B) \quad = \quad \lambda \frac{1}{rank_{EF}(ca,ka)} + (1-\lambda) \frac{1}{rank_{PR}(ca,ka)}$$

## Results

| | #rel | MAP | MRR | P@5 | P@10 | P@20 |
|---|---|---|---|---|---|---|
| EF (baseline) | 576 | 0.196 | 0.531 | 0.336 | **0.332** | **0.269** |
| + Method 1: | | | | | | |
| (A) | 576 | **0.209*** | **0.659*** | **0.396*** | 0.326 | 0.267 |
| (B) $\lambda = 0.5$ | 576 | **0.197** | **0.584*** | **0.376*** | 0.324 | 0.267 |
| + Method 2: | | | | | | |
| (A) | 576 | 0.181 | **0.576*** | **0.340** | 0.292 | 0.242 |
| (B) $\lambda = 0.7$ | 576 | 0.188 | **0.559*** | **0.344** | 0.306 | 0.254 |

- effective in terms of early precision

## Results (2)

| **TREC 2005** | MAP | MRR | P@5 | P@10 | P@20 |
|---|---|---|---|---|---|
| EF | 0.196 | 0.531 | 0.336 | **0.332** | **0.269** |
| EF+EP | **0.209** | **0.659** | **0.396** | 0.326 | 0.267 |

| **TREC 2006** | MAP | MRR | P@5 | P@10 | P@20 |
|---|---|---|---|---|---|
| EF | 0.328 | 0.506 | 0.395 | 0.408 | 0.377 |
| EF+EP | **0.466** | **0.851** | **0.661** | **0.587** | **0.495** |

- TREC'05 topics: effective in terms of early precision
- TREC'06 topics: very effective in all respects

## Outline

- Introduction
- Topical profiles
  - Formal definition
  - Algorithms
  - Evaluation
- An application to expert finding
- Conclusions
- Further work

## Conclusions

- Profiles provide context and evidence to users searching for expertise
- "Inverted" expert finding is not effective
- We proposed two methods and a filtering algorithm that significantly outperformed the baseline
- We applied profiling algorithms to enhance the performance of an expert finding method

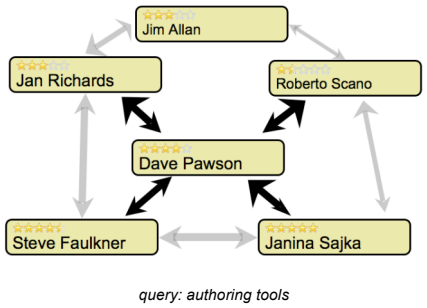
## Further work

- Further investigate the relation between expert finding and profiling
- More sophisticated models for creating profiles (Language Models)
- Evaluating and making use of social profiles
- How do these methods work on a different collection, e.g. university data set?

## Questions

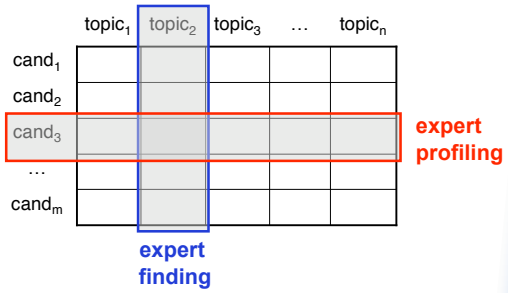
**Krisztian Balog**
kbalog@science.uva.nl

## Social profile




*query: authoring tools*

## Expert finding vs profiling




## Expert finding vs profiling (2)

- Expert finding
  - **p(calq)** *what is the probability of a candidate ca being an expert given the topic q?*
- Expert profiling
  - **p(kalca)** *what is the probability of a knowledge area ka being part of the candidate's profile?*
  - using knowledge area as a query (q=ka) => **p(qlca)**
- Applying Bayes' rule

$$p(ca|q) = \frac{p(q|ca)p(ca)}{p(q)}$$